Briefing Paper

Combined usage statistics as a basis for Research intelligence

Every day millions of interactions take place on the web. Researchers, students, companies and the general public request and read research materials which are available on the internet. These clicks and downloads are being recorded and offer a realm of rich information just waiting to be used. In the present day information infrastructure this information can be collected, exchanged and analysed. Aggregating these interactions can provide data which can inform how research materials are being used. This in turn will provide a much broader insight in the impact research has not only in the research community, but also on business and society.

Business intelligence is a well established approach to collecting information in business as a basis for improved decision making. In the higher education and research sector this concept could be applied to research, the term business intelligence would in that case be translated into research intelligence. This does raise the question which concepts from business intelligence could be applied in the university environment. Reliable numerical data should be available as a basis for decision making. Usage statistics are a clear example of data which can offer a valuable contribution to the information required to make informed decisions.

In the case of usage statistics it is important to distinguish between the raw data of interactions and interpretations of this data. The raw data will need to be translated into meaningful information by means of metrics, the process of interpreting and relating this data. The metrics used in evaluating research are different for different disciplines, i.e. a journal article in medicine will attract a different audience than a monograph in humanities. Having a large set of data available, which can be interpreted by sensible metrics will give room for valuable services. These are the services which would serve the needs of individual researchers, students, management and funders alike (cf. Fig. 1).

Collecting and sharing usage statistics would meet the following aims:

1. It would support decision makers in making more informed decisions. For example in the case of licensed materials the cost per download can be calculated. In the case of Open Access materials the cost of publishing the article can be compared to the use being made of the material. Resisting the sharing of usage data, would result in only have limited information on which to base decisions.

2. It would allow benchmarking, i.e. information would not only be available on a general level, but it would also be possible to compare your results with those from comparable institutions.

3. It would allow for new services and systems based on the combined interpreted data. An example would be a service which will follow students in their behaviour and provide recommendations for further reading, both for students but also for teaching staff.

Information on downloads can offer an interesting insight into user behaviour. At present however there is too little information available to base tools and services on. A richer
collection of context is required to offer meaningful services. An inspiring example are recommender services which would offer a student insight into other relevant research materials which are available. If we wish to explore these possibilities it would be valuable to exchange experiences with companies which have experiences with the collecting of user behaviour. The experiences collected would offer valuable input for research.

To achieve a broad collection of usage statistics collaboration is essential. At present publishers are providing very interesting information environments, yet restricted to their products. By establishing a collaboration mutual benefits could be achieved by both publishers and universities. The higher education sector is not equipped to provide this on its own. From this perspective the lessons learned in the European PEER project¹ can provide interesting insights.

A word of caution should also be expressed here. Statistics can be applied to achieve positive goals however there are dangers. Privacy issues certainly need to be taken into consideration, at present user identity and user behaviour is already being tracked by large commercial entities. To guard the privacy of the users it is important to maintain close connections with the identity provider which offers the user access to the internet and the materials.

Based on the arguments above it would be timely to identify the stakeholders concerned. Probable parties are among others researchers, students, university managers, publishers, funders, society, SMEs, government, libraries and consortia. There are also specific businesses which are already active in this field. We could learn a lot from Google, Amazon, mobile phone companies, supermarkets and the research they have undertaken. This is also a topic being addressed by technical marketeers and research institutes like CWTS² and IFQ³.

A following step would be to identify several possible uses of usage statistics. Services could include recommender services for students and the public alike, services offering feedback to researchers having deposited materials in repositories, but also services to support the creation of communities of persons using selected materials. These services are only very first applications and we expect to find not only intended but also quite unintended and unexpected uses appearing over time. Yet all these services and applications directed at creating a rich source of information to base research intelligence on, will not be possible without a collected and agreed set of usage statistics.

Berlin, 30 March 2009, Frank Scholze, Adrian Price, Natalia Manola, Margit Palzenberger, Daniel Metje, Wilma Mossink, Gerard van Westrienen, Keith Russell

¹ http://www.peerproject.eu/
² http://www.socialsciences.leiden.edu/cwts/
³ http://www.forschungsinfo.de/